统计和词典方法相结合的双语语料库词对齐*

吕雅娟 赵铁军 李生 杨沐昀
（哈尔滨工业大学计算机科学与技术学院 哈尔滨 15001）
Email:{lyj,tjzhao,ymy}@mtlab.hit.edu.cn

摘要： 双语语料库词对齐研究对于自然语言处理的许多应用具有重要意义。本文在对基于统计和基于词典的词对齐方法进行实验分析的基础上，提出了基于词典的语言学信息与词性统计相结合的词对齐方法，该方法在训练语料库规模较小的情况下，充分利用现有资源，取得了较好的词对齐结果。

关键词： 自然语言处理，双语语料库，词对齐

Word Alignment Based on Statistic and Lexicon

Lü Yajuan Zhao TieJun Li Sheng Yang MuYun
Dept. of Computer Science & Engineering , Harbin Institute of Technology, Harbin, 15001
Email:{lyj,tjzhao,ymy}@mtlab.hit.edu.cn

ABSTRACT: Word alignment of bilingual corpus are useful for many NLP applications. On comparing the word alignment results abtained from statistic-based and lexicon-based methods, this paper proposed a hybrid method that combine linguistic information and POS-based statistic. The proposed method achieve good performance based on a small training corpus.

Keywords: NLP, Bilingual Corpus, Word Alignment

1 引言

双语语料库含有两种不同语言之间的对照翻译信息，因此在自然语言处理的许多领域具有很高的研究和使用价值。含有对齐信息（段落、句子、短语、单词级）的双语语料库具有更高的实用价值，它可以为机器翻译、词典编纂、信息检索、翻译知识的获取、词义排歧等自然语言应用提供更大程度的支持。近年来，国内外研究者对于双语语料库的自动对齐技术进行了大量的研究[1-5]。本文在英汉双语句对齐的基础上研究小规模双语语料库上的词汇对齐方法。由于词汇对齐比段落和句子对齐提供了更细粒度的对译信息，因此对于它的研究具有重要意义。

对于词汇级的对齐，目前有两类主要方法：基于统计的方法和基于词典的方法。统计方法的主要思想是通过对大规模双语语料的统计训练，获得双语对译词的同现概率，以此作为对齐的基础。Brown 首先实现了基于统计机器翻译模型的词对齐[6], Dagan, Chang 等对 Brown 的模型进行了改进 [7，8]。Gale 和 Church 等人使用一种类 x^2 的概率分布统计双语

* 本课题得到国家自然科学基金的资助，合同号 69775017

对译词的同现概率[9]。Vogel 等把 HMM 模型引入词对齐[10]。Pascal Fung 提出了一种在非对齐的双语语料中提取部分词汇词对应方法[11]。近年来，基于词典和语言学知识的词对齐方法也逐渐的得到了人们的重视。Ker[12]，王斌[13]等利用语义类进行词对齐。Huang 实现了基于语言学比较的汉朝双语词对齐[14]。另外，孙乐等还利用 Chunk 分析进行汉英双语词汇对齐[15]。

本文在现有资源的基础上首先对基于统计和基于词典的词对齐方法进行了实验和改进，综合分析了各种方法的实验结果。提出了小规模双语语料库上的统计和词典相结合的汉英双语语料库词对齐方法。实验结果表明该方法能够充分利用现有资源，避免了词统计的数据稀疏问题，是小规模双语语料库上有效的词对齐方法。

2 基于词典的词对齐方法

所谓双语词汇对齐，是指在原文本和翻译文本中找到词汇的对译关系。由于双语词典含有词汇的译文信息，因此是进行双语对齐的最有效可靠的资源。我们首先使用英汉双语词典进行了完全基于词典的词对齐实验(DA)。完全基于词典的方法得到的非空词对齐具有较高的正确率，但由于真实翻译中上下文的多样性的和翻译的灵活性，词典译文的覆盖率相对较低。为了提高词典译文的覆盖率，引入了词典的模糊匹配。

2.1 词典模糊匹配

虽然有些对齐词的译文在词典中并没有给出，但是在很多情况下对应译文和词典的译文包含相同的字，例如下面的句对：

The amount of one hundred yuan is written in Chinese capital numeral characters .

100 元人民币要用大写的中国数字填上。

其中，“chinese”的词典译文中有“中国人；中国的；中国话；”，尽管译文中没有“中国”，但是根据词典译文应该也能识别出该对应。因此，这里引入了模糊匹配机制。

汉语词 c_1 和 c_2 的模糊匹配相似度定义为：

$$Sim(c_1,c_2)=\frac{2*|c_1\cap c_2|}{|c_1|+|c_2|}$$

其中，$|c_1\cap c_2|$ 为 c_1 和 c_2 所含的公共字的个数，$|c_1|$ 和 $|c_2|$ 分别为 c_1 和 c_2 所含字数。

英语词 e 和汉语词 c 的匹配相似度为：

$$DTSim(e,c)=\max_{d\in DTe} Sim(d,c)+(Count(Sim(d,c)>h_1)-1)*0.1$$

其中，DTe 为 e 的所有译文。h_1 为相似度的阈值，$Count()$为次数统计函数。

基于模糊匹配的词典对齐(DSimA)，在对齐的正确率下降不大的情况下，召回率有了显著的提高。

2.2 基于语义相似度的对齐

在实验中发现，翻译中常常会有利用同义词替代翻译词的现象。Ker[12]和王[13]在词对齐中都曾引入语义作为词典对齐的补充。这里我们采用文[13]中方法计算语义相似度。

同义词词林是现代汉语比较常用的一部义类词典。它所收词语全部按语义分类编排，共分为 12 大类，94 中类，1428 小类，在词条中用语义代码表示。例如，“成功”的词义代码为 Ie14，表示其处于 I 大类、e 中类、14 小类。可以把整个语义分类体系想象成一棵语义树，根节点的儿子是所有大类，某个大类的儿子是他下属的中类，叶子节点为各个小类。两词义 S_1 与 S_2 之间的语义距离 $SenseDist(S_1,S_2)$ 可以定义为语义树中结点 S_1 到结点 S_2 的最短路径的长度，通过比较两个词的语义编码可计算出它们的语义距离。如，$SenseDist(Aa02,Aa01)=2$，$SenseDist(Ab01,Aa01)=4$，$SenseDist(Ba01,Aa02)=6$。

显然，$SenseDist(S_1,S_2)$ 越小，S_1 与 S_2 在语义上越相似。定义词义 S_1 与 S_2 的语义相似度为：

$$SenseSim(S_1,S_2)=\begin{cases}1/SenseDist(S_1,S_2) & S_1\neq S_2\\ 1 & S_1=S_2\end{cases}$$

在此基础上，定义两个汉语词 c_1、c_2 的语义相似度为：

$$CCClassSim(c_1,c_2)=\max_{\substack{S_m\in Classof(c_1)\\ S_n\in Classof(c_2)}} SenseSim(S_m,S_n)$$

其中，$Classof(S)$ 函数返回词语 S 的词义代码集合。

定义英语词 e 和汉语词 c 的语义相似度为：

$$ECClassSim(e,c)=\max_{d\in DTe} CCClassSim(d,c)$$

实验结果表明，利用语义相似度进行词汇对齐(CSimA)，可以弥补双语词典译文覆盖面的不足。当与基于双语词典的方法(DT+DSimA)相结合时提高了对齐的召回率，但也使对齐的正确率有所下降。

3 基于统计的词对齐方法

尽管基于统计的词对齐方法已被证明有效，但是该方法需要超大规模的双语语料作为训练基础（Brown 所用语料库规模是 1,778,620 句对[6]，Gale 所用语料库的规模是 897,077 句对[9]）。由于目前我们尚难获得如此规模的双语语料库，而在小规模语料库中直接应用这些基于词汇同现的统计方法不可避免出现数据稀疏问题。因此这里对传统的统计方法进行了改进，使其能够适用于小规模语料库的对齐。

3.1 统计方法建立词性对译表

通过对双语语料的研究发现，由于同一词的不同译文往往具有相同或相似的词性类。目前词性标注技术已经比较成熟，我们可以获得比较可靠的双语词性标注，因此可以通过词性的共现实现词的对齐，这样就可以避免数据稀疏问题。这里采用 Dice 系数[12]计算词性之间的对译概率。为了提高统计的精确性，首先用词典方法对语料进行了基于双语词典的对齐(DA)。统计时对于已经得到的对齐仅计算对应词性在该对齐中的同现。统计得到的概率最大的 10 个词性互译对如表 1 所示：

表 1 统计得到的前 10 个概率最大的词性对应及概率

序号	英语词性	汉语词性	互译概率
1	SYM (符号)	sym (符号)	0.800
2	PRP (代词)	r (代词)	0.788
3	MD (情态动词)	vz (助动词)	0.738
4	NN (名词)	ng (名词)	0.645
5	RB (副词)	d（副词）	0.561
6	JJ (形容词)	a (形容词)	0.512
7	VBZ (动词单数第三人称)	vx (系动词)	0.487
8	IN (介词)	p (介词)	0.479
9	ART (冠词)	m (数词)	0.445
10	CD (数词)	m (数词)	0.433

把利用这种方法得到的词性对齐(POSA)与基于双语词典的方法(DT+DSimA)相结合可以大大提高对齐的召回率，并且对齐的正确率也比较高。

3.2 统计方法建立双语对译补充词表

作为双语译文词典的补充，我们用统计方法从语料库中提取了部分新的对译词。采用 Gale 和 Church 提出的联列表法(contingency table)计算词的同现[9]，同现概率用 ϕ^2 来表示：

$$\phi^2 = \frac{(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$

其中：N 表示双语句对总数。

$a = freq(e,c)$: 同时包含英文词 e 和汉语词 c 的句对数

$b = freq(e)\text{-}freq(c,e)$: 仅包含英语词 e 的句对数

$c = freq(c)\text{-}freq(c,e)$: 仅包含汉语词 c 的句对数

$d = N\text{-}a\text{-}b\text{-}c$: 不包含汉语词 c 或英文词 e 的句对数目

ϕ^2 在[0，1]之间，它的值越大则表示这两个词之间的相关程度越高。与词性统计类似这里在统计过程中也考虑了词典对齐结果，不同的是为了发现新译文，排除词典已有译文的影响，在统计时把双语词典已得到的对应排除在统计之外。统计语料库共有双语句对

30,094 对，包含 17156 个英语和 18654 个汉语词汇。其中仅有 2668 个英语词在统计中同现次数大于 1，且对应词各自出现词数大于 2。可以看出，由于统计数据不足，数据稀疏问题比较严重。这里仅把得到的部分高频对译词加入词典译文中。

4 统计和词典相结合的词对齐

通过基于词典和基于统计的词对齐实验，可以看出完全基于词典的对齐可以获得可靠的非空对齐，因此可以作为初始对齐的方法。但是由于双语词典的覆盖面有限，得到的对齐的召回率并不理想。基于语义相似度和基于统计的方法可以弥补纯词典方法的不足，获得更多对齐。因此在现有资源和语料规模的情况下，综合使用基于词典和基于统计的方法可以得到更好的对齐结果。

在对齐的过程中可能出现冲突，这里使用位置相对形变距离和对齐评价函数进行消歧。整个对齐过程通过贪心算法实现。

4.1 位置相对形变距离

位置相对形变距离是一个可能的对齐相对于最近的确定对齐的距离。位置为 i 的英语词对齐为位置为 j 的汉语词的位置相对形变距离定义为：

$$Dis(i,j) = \min(|d_L|,|d_R|)$$
$$d_L = (j - j_L) - (i - i_L)$$
$$d_R = (j - j_R) - (i - i_R)$$

(i_L, j_L) 为左边距离 i 最近的确定对齐的位置对；

(i_R, j_R) 为右边距离 i 最近的确定对齐的位置对；

4.2 对齐评价函数

对于对齐（e, c）的对齐评价函数 $\Pr(e,c)$ 定义为：

$$\Pr(e,c) = T(e,c) * D(e,c)$$

其中 $T(e,c)$为翻译概率，在不同的算法中，翻译概率计算方法不同。如基于模糊词典对齐的方法中 $T(e,c)$=DTSim(e,c)，在基于词性统计的对齐中 $T(e,c)$=POSSim(pos(e),pos(c))，在词典与词性统计相结合的对齐中 $T(e,c)$= DTSim(e,c)+ POSSim(pos(e),pos(c))，以此类推。

$D(e,c)$为位置形变概率，取值如下：

$$D(e,c) = \begin{cases} d_1 & if \quad Dis(i,j) = 0; \\ d_2 & if \quad Dis(i,j) = 1,2; \\ d_3 & if \quad Dis(i,j) = 3,4; \\ d_4 & if \quad Dis(i,j) > 4; \end{cases}$$

其中 d_1,d_2,d_3,d_4 的值通过标准对齐语料统计得到。d_1 =0.226, d_2 =0.043，d_3 =0.017，d_4 =0.004。

4.3 基于贪心算法的对齐过程

整个词对齐过程和对齐歧义消除过程通过贪心算法实现。步骤如下：

1. 建立完全基于词典的候选对齐
2. 建立可靠连接。对于词典对齐中得到的无冲突对齐建立连接关系，作为初始可靠连接。另外对于英语和汉语各增加一个首节点和尾节点，建立对应连接作为可靠连接。
3. 根据算法需要建立词典模糊对齐，语义对齐或统计词性对齐等候选对齐，并计算相对形变概率和对齐评价函数。
4. 选择大于一定阈值的具有最大概率的候选对齐，建立该对齐；
5. 删除与选定对齐有冲突的对齐；
6. 重新计算候选对齐的位置形变概率和对齐评价函数
7. 重复（4），（5），（6），直到没有合乎要求的候选对齐

5 实验结果及分析

实验中使用的英汉双语词典包含词条 116,757 条，其中有短语 21,741 条。语义词典使用《同义词词林》，含有 53,146 条汉语词条。统计训练用的经过句子对齐的双语句对 30,094 对，来自初高中英语和大学英语课本，其中的汉语句子经过分词处理。从训练语料中随机抽取 10,000 句对并进行词对齐的人工校对，作为标准测试语料。词对齐结果用正确率和召回率进行评价，正确率和召回率定义如下：

$$\text{正确率} = \frac{\text{正确对齐的词对数}}{\text{得到的对齐词对总数}} \times 100\%$$

$$\text{召回率} = \frac{\text{正确对齐的词对数}}{\text{标准对齐文本中的词对总数}} \times 100\%$$

表 2，表 3 分别给出了基于词典方法和基于混合方法在给定阀值下的词对齐结果。根据评价时是否包含空对齐（双语句对中一种语言中的词在另一种语言中没有对齐词），召回率和正确率的计算也分成两种情况，表 2 中为非空对齐评价结果。表 3 给出了对比结果。

从表 2 可以看出基于双语词典的对齐(DA)非空匹配具有较高的正确率，但由于词典的覆盖能力有限，因此召回率较低。词典模糊匹配(DSimA)在基本保证正确率的情况下，可以提高非空匹配的召回率。基于语义词典的对齐(CSimA)尽管单独使用时正确率和召回率都不理想，但与基于双语词典相结合时可以使召回率有较大幅度的提高，但是正确率也略有下降。

从表 3 可以看出，尽管基于双语词典的对齐具有很高的非空对齐正确率，但由于词典

的覆盖率较低，因此当考虑空对齐词时整体的正确率和召回率都不高。基于词性统计的方法（POSA）单独不能获得可靠的正确率，因此不能作为独立的对齐方法。把基于词典和基于词性统计的方法相结合既可以保证非空对齐的正确率，又使召回率有了大幅度的提高。这时在包含空对齐的情况下召回率和正确率也都达到了几个实验的最佳值。当综合所有方法（包括基于词典的对齐，基于语义词典的对齐和基于词性统计的对齐）进行对齐实验时，其结果（表 3 最后一行）反而不如只用双语词典和词性统计相结合的方法。通过分析发现这是因为当得到的候选对齐太多时，对齐冲突也相应增加，当消歧出现错误时正确率反而降低。

表 2　基于词典方法的词对齐结果

对齐方法		对齐数	正确对齐数	正确率(%)	召回率(%)
DA		41487	38783	93.48	48.76
DSimA	>0.5	49962	45820	91.71	57.61
	>0.7	53007	49185	92.79	61.84
CSimA	≥0.5	50293	24706	49.18	31.06
	≥1	41619	29254	70.29	36.78
DA+DSimA(>0.7)+CSimA(>0.5)		68396	60489	88.44	76.05

表 3　基于混合方法的词对齐结果

对齐方法	非空对齐		包含空对齐	
	正确率(%)	召回率(%)	正确率(%)	召回率(%)
DA	93.48	48.76	68.96	74.38
POSA(>0.2)	64.49	28.36	41.27	49.46
DA+DSimA(>0.5)+POSA(>0.2)	90.16	76.37	80.87	78.75
DA+DSimA(>0.5)+CsimA(>0.5)+POSA(>0.2)	87.08	77.48	80.74	76.47

分析对齐中的错误，一部分(<30%)是由于资源不足引起的(词典译文缺乏，统计数据不足等)。另外有大约 30%的错误是由于短语匹配造成的。此外的错误大部分是由于汉语和英语之间存在固有的表达差异造成的。如汉语中的成语，谚语，惯用搭配在相应的英文中通常采用意译，另外汉语表达中的一些特殊现象，如离合词，重叠词等，也是引起一些错误的原因。本文提到词对齐方法尚不能解决后两类错误，对于这些错误的解决有待进一步增加句法分析和语言学知识加以解决。

6 结论

通过对基于词典和基于统计的词对齐方法的实验分析，可以得到以下结论：

a) 语言学信息在双语语料库词对齐中有着重要作用。双语词典可以提供可靠的非空对齐。基于语义相似度的方法可以提高对齐的召回率。

b) 当语料库规模较小时，基于词性的统计方法对提高对齐的召回率具有重要作用，

c) 在资源和语料不足的情况下，基于词典和基于词性统计相结合的方法是进行词对齐的有效方法。

尽管本文使用了多种对齐方法，但对齐的召回率仍然不能令人满意。一个主要原因是由于汉英双语间的语言差异，使得很多对齐情况需要在短语分析的层面上才能得以解决。下一步工作我们将以现有的词对齐为基础实现短语级的对齐，并在此基础上实现翻译知识的自动获取。

参考文献

[1] K. W. Church. "Char-align: a Program for Aligning Parallel Texts at the Character Level." Proc. of the 31st Annual Meeting of the ACL. 1993: 1-8

[2] P. F. Brown, J. C. Lai and R. L. Mercer. "Aligning Sentences in Parallel Corpora." ACL-91:169-176

[3] T. Utsuro. etc. " Aligning a Parallel English-Chinese Corpus Statistically with Lexical Criteria." ACL-1994: 80-87.

[4] H. Wantanabe, S. Kurohashi, and Eiji Aramak. "Finding Structural Correspondences from Bilingual Parsed Corpus for Corpus-based Translaton". Coling-2000.

[5] 刘昕，周明，朱胜火，黄长宁. 基于自动抽取词汇信息的双语句子对齐.计算机学报. 1998, 21(8):151-158

[6] P. F. Brown. ect. "The Mathematics of Statistical Machine Translation: Parameter Estimation" Computational Linguistics, Vol 19, No.2 ,1993

[7] Dagan. I., Church K.etc. "Robust Bilingual Word Alignment for Machine Aided Translation" Proceedings of the Wordkshop on Very Large Corpora: Academic and Indestrial Perspectives,pages 1-8, Columbus,1993.

[8] Chang, J.S., and Chen, M.H.C. "Using PartialAligned Parallel Text and Part-of-speech Information in Word Alignment." In Proceedings of the First Conference of the Association for Machine Translation in the Americas(ATMA'94), pp16-23.

[9] W. A. Gale, K. W. Church "Identifying Word Correspondences in Parallel Texts." Proceedings of the 4th DARPA workshop on Speech and Natual Language. 1999:152-157

[10] Stephan Vogel, Hermann Ney, Christoph Tillmann. "HMM-based Word Alignment in Statistical Translation." Coling-96

[11] Pascale Fung. "A Statistical View on Bilingual Lexicon Extraction: from Parallel Corpora to Non-parallel Corpora." Lecture Notes in Artificial Intelligence, Springer Publisher, vol 1529,1998.

[12] Sue J. Ker and Jason S. Chang. "A Class-based Approach to Word Alignment" Computational Linguistics 23(2):313-343,1997.

[13] 王斌，刘群，张祥. "汉英双语库词汇对齐研究." 计算语言学文集，清华大学出版社，1999.

[14] Jin-Xia Huang, Key-Sun Choi. "Chinese-Korean Word Alignment Based on Linguistic Comparison. ". ACL-2000

[15] Sun Le, Jin Youbing, Du Lin, Sun Yufang. "Word Alignment of English-Chinese Bilingual Corpus Based on Chunks." Proc. Of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. P110-116.